觀察資料
確定分析方向

1. 合併 2019 Oct-2020 Feb 共五個月之原始資料：
   - 資料量：20,692,840 rows * 10 columns (下圖顯示前三列)

	event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
0	2019-10-01 00:00:00 UTC	cart	5773203	1487580005134238553	NaN	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
1	2019-10-01 00:00:03 UTC	cart	5773353	1487580005134238553	NaN	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
2	2019-10-01 00:00:07 UTC	cart	5881589	2151191071051219817	NaN	lovely	13.48	429681830	49e8d843-adf3-428b-a2c3-fe8bc6a307c9

2. 觀察資料完整性：

	column_name	percent_missing
event_time	event_time	0.000000
event_type	event_type	0.000000
product_id	product_id	0.000000
category_id	category_id	0.000000
category_code	category_code	98.291225
brand	brand	42.319551
price	price	0.000000
user_id	user_id	0.000000
user_session	user_session	0.022220

基於：

- category_code 與 brand 缺值過多，故不採用
- 無 user 年齡層、性別等其他資訊

分析方向：

- 採用 RFM* 分析模型進行顧客分群

RFM*
- Recency 最近一次消費日期
- Frequency 消費頻率
- Monetary 總消費金額

資料前處理

1.排除該次分析不需要的欄位：

- 僅留下四欄位：event_time, event_type, price, user_id
- 資料量：20,692,840 rows * 4 columns

	event_time	event_type	price	user_id
0	2019-10-01 00:00:00 UTC	cart	2.62	463240011
1	2019-10-01 00:00:03 UTC	cart	2.62	463240011
2	2019-10-01 00:00:07 UTC	cart	13.48	429681830

2. 資料補值/清理：

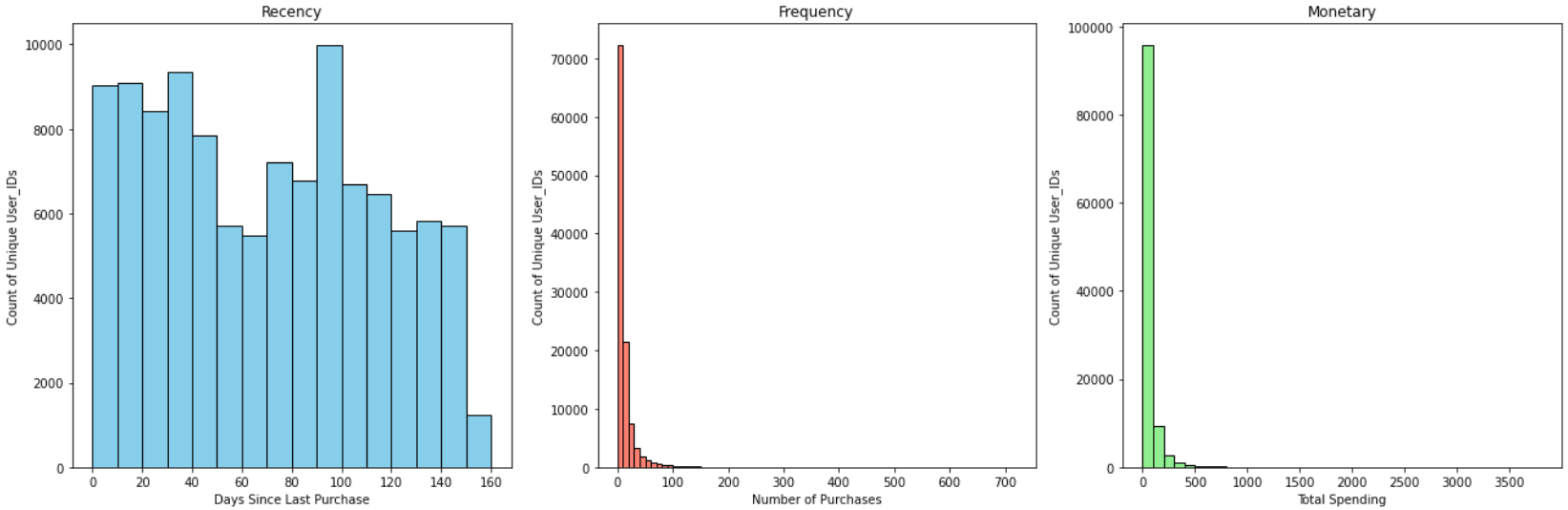
- 上述四個欄位皆已無空值
- Duplicates: 經檢測有 1,629,333 筆重複資料，然考慮可能為一秒內連按兩次新增/移除，故不刪除（推測為有效動作）
- 新增 ‘month’ 欄位，以利後續分析觀察各月趨勢

RFM分析前處理 (1)

1. 建立 RFM Metrics (右圖例):
   - Recency: 上次消費日離 2020-03-01 距離天數
     (資料截止日為2020-02-29)
   - Frequency: 總消費次數
   - Monetary: 總消費金額

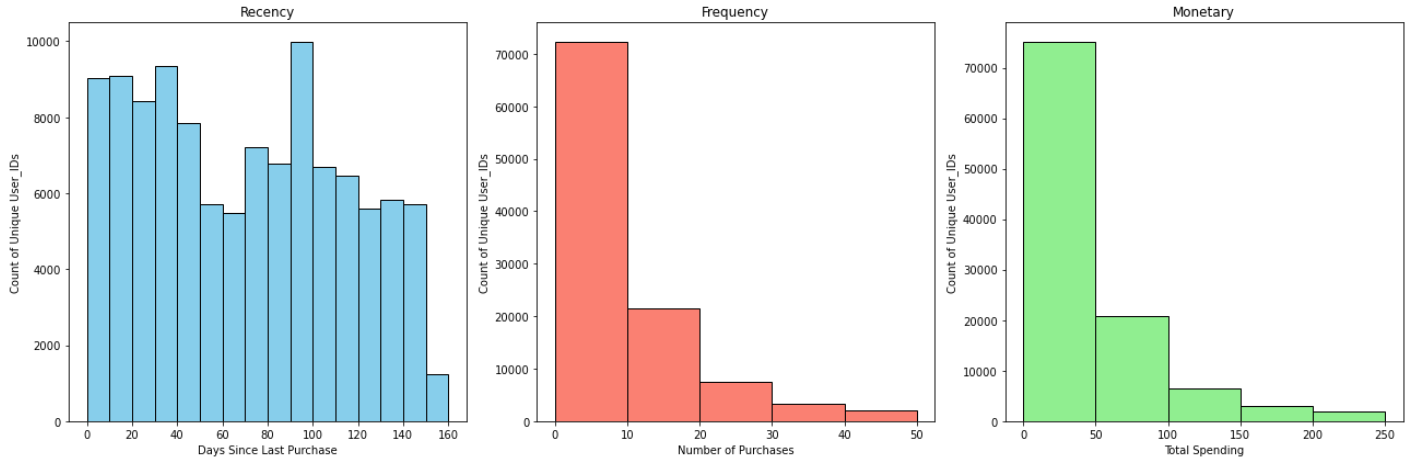
user_id	recency	frequency	monetary
9794320	96	4	12.68
10079204	115	2	25.81
10280338	10	86	177.83

2. 觀察資料分布：
   - 可見 Frequency, Monetary 分佈呈現嚴重右偏態，故決定進一步處理

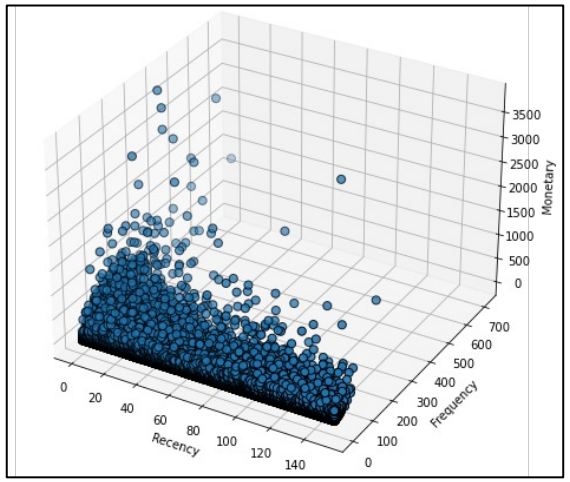


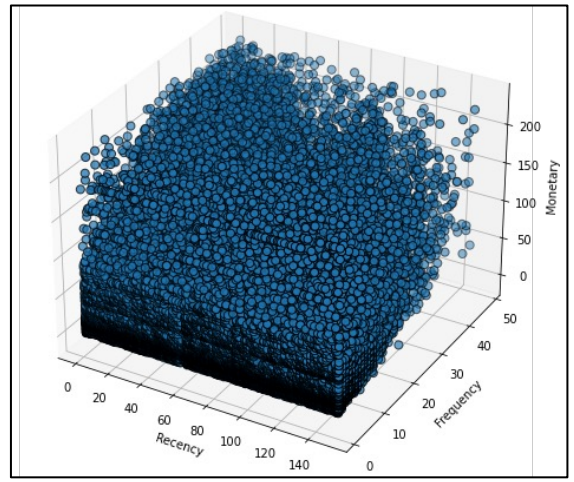
RFM分析前處理(2)

1. 排除 RFM Metrics 離群值：
   - 分別將 R, F, M 中超過兩倍標準差者視為 outliers 並排除，經處理後結果如下



補充：三維分佈之比較圖 (before v.s. after)

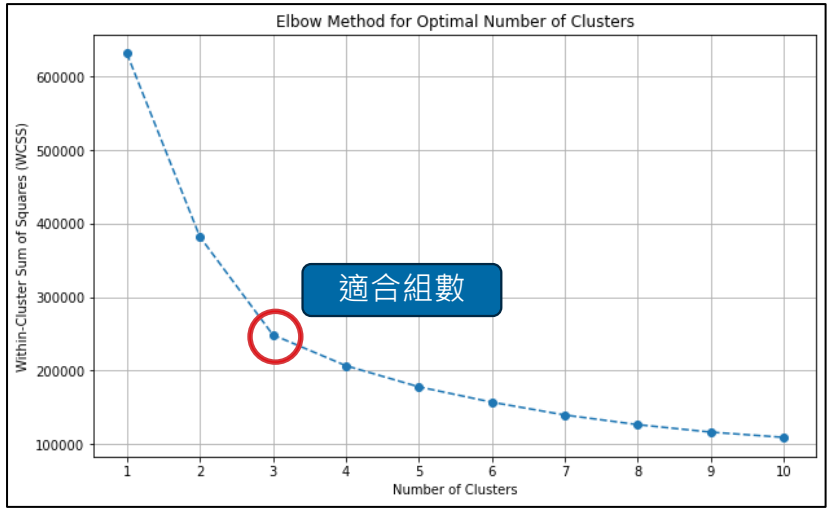


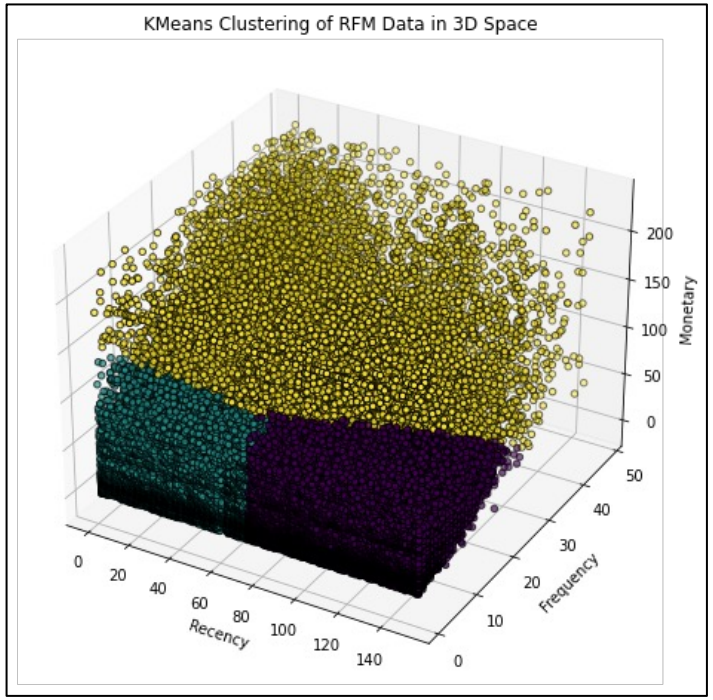


RFM分析
K-Means 分群

嘗試用 K-Means 將 users 分群：

1. 透過 elbow method 找出適合組數（左圖）
2. 將分群後資料繪出（右圖）





K-Means 將 users 主要分為：

1. 紫：消費頻率低、近期不活躍、消費金額中低
2. 綠：消費頻率低、近期有消費、消費金額中低
3. 黃：消費頻率高、消費金額中高

消費力強

消費力弱

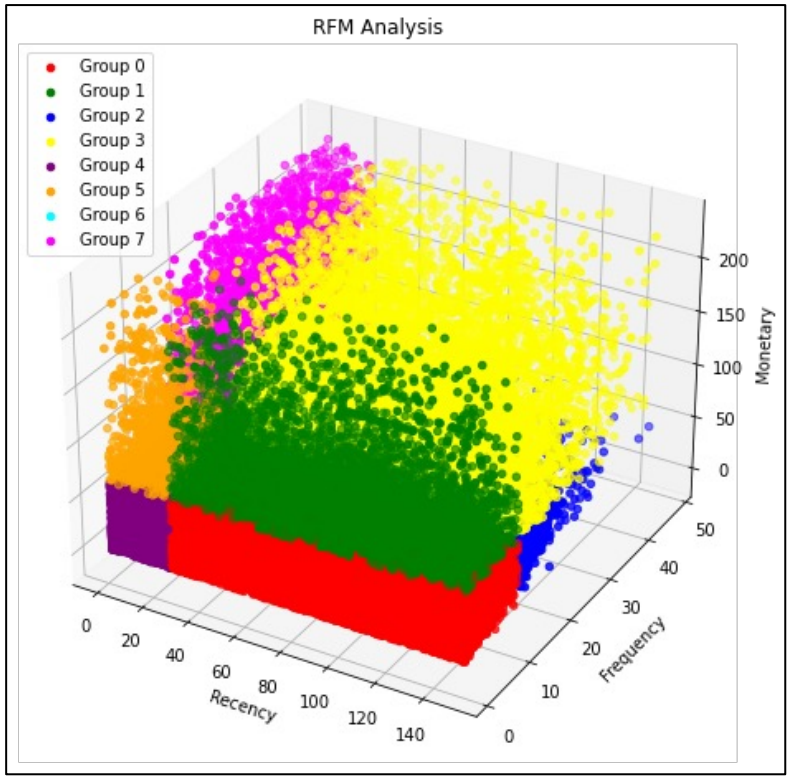
RFM分析
80-20 分群 (1)

因 K-Means 分出之3群稍嫌粗略：

- 採用傳統統計方法，分別將 R, F, M 切分為 top 20% & last 80%

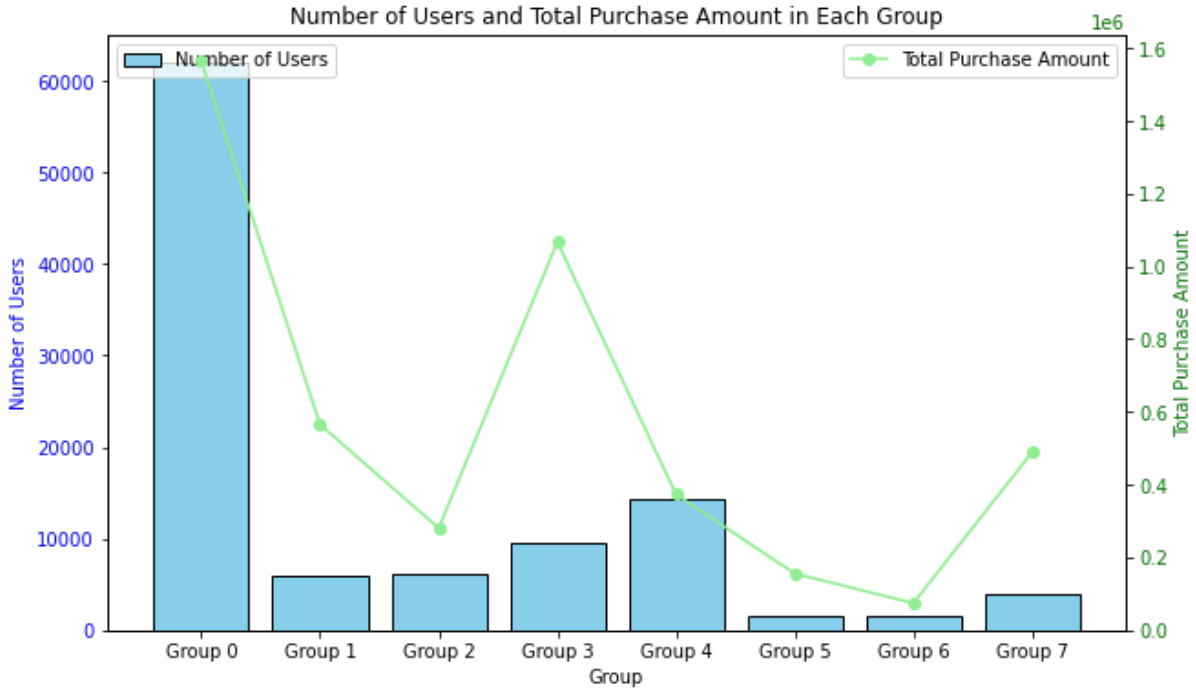
以 80-20 分出之 8 群分別為：

- G0: 遠R, 低F, 低M → 一般舊客
- G1: 遠R, 低F, 高M → 重要開發客
- G2: 遠R, 高F, 低M → 一般舊常客
- G3: 遠R, 高F, 高M → 重要挽留舊客
- G4: 近R, 低F, 低M → 一般新客
- G5: 近R, 低F, 高M → 重要新客
- G6: 近R, 高F, 低M → 一般常客
- G7: 近R, 高F, 高M → VIP 客戶



RFM分析
80-20 分群 (2)

下圖顯示 8 組之「用戶數量」與「消費總額」：



- 其中消費金額部分以：G0 (一般舊客), G3 (重要挽留舊客), G1 (重要開發客), G4 (一般新客) 為大部分收入來源
- 用戶數聚集於 G0 (一般舊客) 居多，除此則以 G4 (一般新客), G3 (重要挽留舊客) 較為明顯

RFM分析
經營建議

針對主要族群建議：

- 針對 G0 (一般舊客)：該客群曾來訪過平台，但沒有成功建立長期關係，可以再度推播信件或廣告再度曝光，但恐無需針對性地投入大量成本。
- 針對 G3 (重要挽留舊客)：該客群曾高頻且大量地在該平台消費，但近期不活躍。應主動聯繫了解長期未消費之原因，並加以解決，避免該客群流失。
- 針對 G1 (重要開發客)：該客群有高消費力，卻缺乏對平台之忠臣度。應主動提供誘因，建立與該客群之長期關係。
- 針對 G4 (一般新客)：應為首次接觸該平台之一般消費者，可透過觀察該客群了解一般用戶的二訪率，近一步優化平台。

其他建議：

- 因維護舊客相較開發新客容易，且舊客更願在新商品上花費與嘗試，應試圖將 RMF 結構盡量導向舊客為主的結構。

附錄
其他分析

Add-to-Cart Conversion Rate：

- 針對電商平台，購物車轉換率也是一項重要指標
- 依照 OBERLO 2023 電商調查，化妝保養品的轉換率落於 12.18%左右
- 而該資料的平均購物車轉換率為 22.31%，算是合格的表現！